

WEB-BASED CUSTOMER PROSPECTS HARVESTER SYSTEM

~~RELATED PATENT APPLICATIONS~~

~~This application claims the benefit of U.S. Provisional Application No. 60/206,772, filed May 25, 2000 and entitled "Server Log File System Utilizing Text Mining Methodologies and Technologies". The present patent application and additionally the following patent applications are each conversions from the foregoing provisional filing: Patent Application Serial No.~~

~~\_\_\_\_\_ (Attorney Docket No. 068082.0105) entitled "Web-Based Customer Lead Generator System" and filed May 21, 2001; Patent Application Serial No. \_\_\_\_\_ (Attorney Docket No. 068082.0111) entitled "Database Server System for Web-Based Business Intelligence" and filed \_\_\_\_\_; Patent Application Serial No.~~

~~\_\_\_\_\_ (Attorney Docket No. 068082.0112) entitled "Data Mining System for Web-Based Business Intelligence" and filed \_\_\_\_\_; Patent Application Serial No.~~

~~\_\_\_\_\_ (Attorney Docket No. 068082.0113) entitled "Text Mining System for Web-Based Business Intelligence" and filed \_\_\_\_\_; Patent Application Serial No.~~

~~\_\_\_\_\_ (Attorney Docket No. 068082.0115) entitled "Text Indexing System for Web-Based Business Intelligence" and filed \_\_\_\_\_; Patent~~

~~Application Serial No. \_\_\_\_\_ (Attorney Docket No. 068082.0116) entitled "Digital Voice Recording Mining System for Web-Based Business Intelligence" and filed \_\_\_\_\_.~~

TECHNICAL FIELD OF THE INVENTION

This invention relates to electronic commerce, and more particularly to a method of acquiring leads for prospective customers, using Internet data sources.

09862314.052101

BACKGROUND OF THE INVENTION

Most small and medium sized companies face similar challenges in developing successful marketing and sales campaigns. These challenges include locating qualified prospects who are making immediate buying decisions. It is desirable to personalize marketing and sales information to match those prospects, and to deliver the marketing and sales information in a timely and compelling manner. Other challenges are to assess current customers to determine which customer profile produces the highest net revenue, then to use those profiles to maximize prospecting results. Further challenges are to monitor the sales cycle for opportunities and inefficiencies, and to relate those findings to net revenue numbers.

Today's corporations are experiencing exponential growth to the extent that the volume and variety of business information collected and accumulated is overwhelming. Further, this information is found in disparate locations and formats. Finally, even if the individual data bases and information sources are successfully tapped, the output and reports may be little more than spreadsheets, pie charts and bar charts that do not directly relate the exposed business intelligence to the companies' processes, expenses, and to its net revenues.

INS. BI With the growth of the Internet, one trend in developing marketing and sales campaigns is to gather customer information by accessing Internet data sources. Internet data intelligence and data mining products face specific challenges. First, they tend to be designed for use by technicians, and are not flexible or intuitive in

ATTORNEY'S DOCKET  
068082.0114

PATENT APPLICATION

4

their operation; secondly, the technologies behind the  
various engines are changing rapidly to take advantage of  
advances in hardware and software; and finally, the  
results of their harvesting and mining are not typically  
5 related to a specific department goals and objectives.

068082.0114

SUMMARY OF THE INVENTION

One aspect of the invention is a web-based computer system for providing, to a business enterprise client, customer lead information from Internet sources.

- 5 Overall, the system may be described as an application service system, having a crawler process that retrieves specified Internet web site data, and a web archive for storing the unstructured data. A harvester process is programmed to accept client criteria describing business prospects and their attributes, to search unstructured  
10 Internet data for prospects matching those criteria and their attributes, and to deliver the results of the search to the client with a link to a document that verifies the prospect's match to the criteria. As with  
15 conventional application service systems, it is accessible by client browser systems via the Internet.

096314.052101

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 illustrates the operating environment for a web based lead generator system in accordance with the invention.

5       FIGURE 2 illustrates the various functional elements of the lead generator system.

FIGURE 3 illustrates a first embodiment of the prospects harvester.

10       FIGURE 4 illustrates a second embodiment of the prospects harvester.

FIGURE 5 illustrates the security features of the lead generator system.

068082.0114

DETAILED DESCRIPTION OF THE INVENTION

Lead Generator System Overview

FIGURE 1 illustrates the operating environment for a web-based customer lead generation system 10 in accordance with the invention. System 10 is in communication, via the Internet, with unstructured data sources 11, an administrator 12, client systems 13, reverse look-up sources 14, and client applications 15.

10 The users of system 10 may be any business entity that desires to conduct more effective marketing campaigns. These users may be direct marketers who wish to maximizing the effectiveness of direct sales calls, or e-commerce web site who wish to build audiences.

In general, system 10 may be described as a web-based Application Service Provider (ASP) data collection tool. The general purpose of system 10 is to analyze a client's marketing and sales cycle in order to reveal inefficiencies and opportunities, then to relate those discoveries to net revenue estimates. Part of the latter process is proactively harvesting prequalified leads from external and internal data sources. As explained below, system 10 implements an automated process of vertical industry intelligence building that involves automated reverse lookup of contact information using an email address and key phrase highlighting based on business rules and search criteria.

More specifically, system 10 performs the following tasks:

- Uses client-provided criteria to search Internet postings for prospects who are discussing products or services that are related to the client's business offerings
- 5 • Selects those prospects matching the client's criteria
- Pushes the harvested prospect contact information to the client, with a link to the original document that verifies the prospects interest
- Automatically opens or generates personalized sales  
10 scripts and direct marketing materials that appeal to the prospects' stated or implied interests
- Examines internal sales and marketing materials, and by applying data and text mining analytical tools, generates profiles of the client's most profitable  
15 customers
- Cross-references and matches the customer profiles with harvested leads to facilitate more efficient harvesting and sales presentations
- In the audience building environment, requests  
20 permission to contact the prospect to offer discounts on services or products that are directly or indirectly related to the conversation topic, or to direct the prospect to a commerce source.

System 10 provides open access to its web site. A  
25 firewall (not shown) is used to prevent access to client records and the entire database server. Further details of system security are discussed below in connection with FIGURE 5.

Consistent with the ASP architecture of system 10,  
30 interactions between client system 13 and system 10 will typically be by means of Internet access, such as by a web portal. Authorized client personnel will be able to



create and modify profiles that will be used to search designated web sites and other selected sources for relevant prospects.

Client system 11 may be any computer station or  
5 network of computers having data communication to lead generator system 10. Each client system 11 is programmed such that each client has the following capabilities: a master user account and multiple sub user accounts, a user activity log in the system database, the ability to  
10 customize and personalize the workspace; configurable, tiered user access; online signup, configuration and modification, sales territory configuration and representation, goals and target establishment, and online reporting comparing goals to target (e.g.,  
15 expense/revenue; budget/actual).

34 Administration system 14 performs such tasks as  
account activation, security administration, performance  
monitoring and reporting, assignment of master userid and  
licensing limits (user seats, access, etc.), billing  
20 limits and profile, account termination and lockout, and  
a help system and client communication.

System 10 interfaces with various client  
applications 15. For example, system 10 may interface  
with commercially available enterprise resource planning  
25 (ERP), sales force automation (SFA), call center, e-commerce, data warehousing, and custom and legacy applications.

09362814.052101

Lead Generator System Architecture

FIGURE 2 illustrates the various functional elements of lead generator system 10. In the embodiment of FIGURE 2, the above described functions of system 10 are  
5 partitioned between two distinct processes.

A prospects harvester process 21 uses a combination of external data sources, client internal data sources and user-parameter extraction interfaces, in conjunction with a search, recognition and retrieval system, to  
10 harvest contact information from the web and return it to a staging data base 22. In general, process 21 collects business intelligence data from both inside the client's organization and outside the organization. The information collected can be either structured data as in  
15 corporate databases/spreadsheet files or unstructured data as in textual files.

Process 21 may be further programmed to validate and enhance the data, utilizing a system of lookup, reverse lookup and comparative methodologies that maximize the  
20 value of the contact information. Process 21 may be used to elicit the prospect's permission to be contacted. The prospect's name and email address are linked to and delivered with ancillary information to facilitate both a more efficient sales call and a tailored e-commerce sales  
25 process. The related information may include the prospect's email address, Web site address and other contact information. In addition, prospects are linked to timely documents on the Internet that verify and highlight the reason(s) that they are in fact a viable  
30 prospect. For example, process 21 may link the contact data, via the Internet, to a related document wherein the contact's comments and questions verify the high level

value of the contact to the user of this system (the client).

A profiles generation process 25 analyzes the user's in-house files and records related to the user's existing customers to identify and group those customers into profile categories based on the customer's buying patterns and purchasing volumes. The patterns and purchasing volumes of the existing customers are overlaid on the salient contact information previously harvested to allow the aggregation of the revenue-based leads into prioritized demand generation sets. Process 25 uses an analysis engine and both data and text mining engines to mine a company's internal client records, digital voice records, accounting records, contact management information and other internal files. It creates a profile of the most profitable customers, reveals additional prospecting opportunities, and enables sales cycle improvements. Profiles include items such as purchasing criteria, buying cycles and trends, cross-selling and up-selling opportunities, and effort to expense/revenue correlations. The resulting profiles are then overlaid on the data obtained by process 21 to facilitate more accurate revenue projections and to enhance the sales and marketing process. The client may add certain value judgments (rankings) in a table that is linked to a unique lead id that can subsequently be analyzed by data mining or OLAP analytical tools. The results are stored in the deliverable database 24.

#### Data Sources

FIGURE 3 provides additional detail of the data sources of FIGURES 1 and 2. Access to data sources may

be provided by various text mining tools, such as by the crawler process 31 or 41 of FIGURES 3 and 4.

INS. 35  
One data source is newsgroups, such as USENET. To access discussion documents from USENET newsgroups such as "news.giganews.com", NNTP protocol is used by the crawler process to talk to USENET news server such as "news.giganews.com." Most of the news servers only archive news articles for a limited period (giganews.com archives news articles for two weeks), it is necessary for the iNet Crawler to incrementally download and archive these newsgroups periodically in a scheduled sequence. This aspect of crawler process 31 is controlled by user-specified parameters such as news server name, IP address, newsgroup name and download frequency, etc.

Another data source is web-Based discussion forums. The crawler process follows the hyper links on a web-based discussion forum, traverse these links to user or design specified depths and subsequently access and retrieve discussion documents. Unless the discussion documents are archived historically on the web site, the crawler process will download and archive a copy for each of the individual documents in a file repository. If the discussion forum is membership-based, the crawler process will act on behalf of the authorized user to logon to the site automatically in order to retrieve documents. This function of the crawler process is controlled by user specified parameters such as a discussion forum's URL, starting page, the number of traversal levels and crawling frequency.

A third data source is Internet-based or facilitated mailing lists wherein individuals send to a centralized

location emails that are then viewed and/or responded to by members of a particular group. Once a suitable list has been identified a subscription request is initiated. Once approved, these emails are sent to a mail server where they are downloaded, stored in system 10 and then processed in a fashion similar to documents harvested from other sources. The system stores in a database the filters, original URL and approval information to ensure only authorized messages are actually processed by system

10 10.

INS 36  
A fourth data source is corporations' internal documents. These internal documents may include sales notes, customer support notes and knowledge base. The crawler process accesses corporations' internal documents from their Intranet through Unix/Windows file system or alternately be able to access their internal documents by riding in the databases through an ODBC connection. If internal documents are password-protected, crawler process 31 acts on behalf of the authorized user to logon to the file systems or databases and be able to subsequently retrieve documents. This function of the crawler process is controlled by user-specified parameters such as directory path and database ODBC path, starting file id and ending file id, and access frequency. Other internal sources are customer information, sales records, accounting records, and call center digital voice records.

A fifth data source is web pages from Internet web sites. This function of the crawler process is similar to the functionality associated with web-discussion-forums. Searches are controlled by user-specified

parameters such as web site URL, starting page, the number of traversal levels and crawling frequency.

Prospects Harvesting From External and Internal Data Sources

5

Referring to FIGURE 3, the prospects harvester process 21 of system 10 may be implemented so as to mine data from both internal and external sources.

Crawler process 31 is a background process (hourly, daily or weekly), operating on any of the sources described above. It performs an incremental newsgroup download and an incremental and traversal web page download. It may provide a robust interface with text fields in relational databases. Crawler process 31 operates in response to use input that specifies a particular web site or sites. Once downloaded, the Internet data is stored in a database 32.

Crawler process 31 may also be used for permission-based email. The crawler technology is applied to identify and extract emails. It applies marketing and business rules to generate email text that elicits permission from prospect. It may pass filtered and opt-in emails to client. This process may be automatically generated or generated manually by the client.

A harvester process 33 provides extraction of contact information from database 32, based on search criteria. Additional features are a thesaurus/synonym search, automatic email cleansing (remove standard "no spam" and distracter characters), comprehensive reverse lookup of value-add business information, and keyword-based sales prospects prioritizing.

A value-add process 34 provides robust and mandatory lead ranking and tracking functionality, operating either

on-line or off-line. It reports basic customer and prospect profiling (i.e., purchasing criteria, time to purchase, pricing window or sensitivity). It may export and import from/to third party sales management or contact management system. It provides search and sub search based on keywords and business criteria, configurable synonym search (add/delete/modify list of related/similar word searches). It may prioritize leads based on keywords and business criteria. It reports potential revenue projections, user and management reporting of lead tracking (new, open, closed, results). It may perform auto email authoring that incorporates intelligent information from prospect's web document and internal business rules. It may further provide an enhanced document summary that contains a short synopsis of the web-based document's context.

A reverse look-up process 35 implements a cascade, multi-site web search for contact information on email addresses. It may search and parse a document for information, to include vcf-type data. It may use a standard reverse email lookup. It may perform a web site search, when email can be linked to a valid company/business URL. It may further parse an email address into name to be used in online white or yellow pages search. It is an intelligent process that eliminates obvious incorrect contacts. For example, if it is known that the contact is from Texas, eliminate all contacts that are not from that state/location.

#### Prospects Harvesting From External Data Sources

FIGURE 4 illustrates another implementation of the prospects harvesting process 21 of FIGURE 2.

Crawler process 41 collects information and documents from the Internet. It archives these documents collected from different sources whenever necessary to keep a historical view of the business intelligence.

5           Indexer 42 indexes the documents retrieved by the crawler 41 and provides the interface for the client to perform searches and sub-searches on specific sets of documents. It also facilitates (1) document keyword highlighting, (2) the extraction of key phrases from  
10 documents and (3) subsequently generates a summary from those documents. ThemeScape, UdmSearch or similar packages may be used to index, search and present documents. Indexer process 42 provides support for multiple file formats such as HTML, DHTML, plain text  
15 (ASCII), Word document, RTF and relational database text fields. Indexer process 42 can either interact with crawler process 41 or access web file archives directly to retrieve documents in different formats (such as Text, HTML and Doc formats). These documents are then indexed  
20 and categorized with their keywords and/or key phrases, date of creation, a brief summary of the original documents and links to the original documents. Links may be either URLs, file path or a path to a database field. This indexing process will be performed on an ongoing  
25 basis as discussion articles and web pages are incrementally downloaded. The results are stored in a central location in the database for future access.

          Harvester process 43 queries the index database 42a using user input keywords, default buyer phrases,  
30 synonyms related to the keywords and predefined stop words. The end results of this process are a set of documents linked to the original documents with



preliminary ranking based on keyword relevance.

Harvester process 43 then follows these links to extract an email address, telephone number and other contact information from the original documents, either through  
5 file archives or web pages on the Internet. The latter functions are based on a set of keywords and parameters specified by customers. The resulting information is then subsequently indexed and cleansed. These email addresses are then entered into a relational database  
10 that is cross-correlated with keywords, source, time stamp, demographics information and company profile information. The harvesting results may be organized and stored into the prospects database 43a with contact information, original document links and preliminary  
15 rankings.

A value-add process 44 adds robust business intelligence to the harvesting process by linking sales prospects with comprehensive and updated business profile information (such as industry, company size, company  
20 business focus and company purchasing budget). Key aspects of this value-add service is accomplished through partnerships with business information sources, such as Harte-Hanks, Hoovers and Dunn & Bradstreet. Reverse lookups may be performed against these business  
25 information sources. Combined with harvested business intelligence, this additional business profile information allows organizations to utilize personalized conversations with prospects, thus dramatically improving their sales close ratios and reducing the time and effort  
30 required to close the sale. The overall ranking of a sales prospect is based on the prospect's business profile, and the keyword relevance in harvested

documents. Using a ranking algorithm, highly targeted and highly qualified sales/marketing prospects may be identified.

5 A mailer process 45 provides an auto-scripting utility for sales people to store pieces of their sales scripts in a knowledge base system. Once stored in the knowledge base they can be copied and pasted into a sales correspondence or used by an auto scripting tool 45a to generate sales correspondence on-the-fly based on the discussion context associated with sales leads. The mailer process 45 provides opt-in/opt-out interface 45b to the harvesting database. When the prospects receive a promotion or other sales correspondence, they will be given the choice to opt-in or opt-out from the lead system if they are not interested in receiving further information.

#### Security

FIGURE 5 illustrates system security. For the security framework to work effectively, the following assumptions are made: database servers exist behind a properly configured firewall, the web server is located outside the firewall in order for users to be able to access the site and login, the application servers exist behind the same firewall, the only allowed traffic from outside the firewall is to the HTTP and HTTPS ports of the application servers. No other access is permitted.

The task of protecting the application servers and the database servers from unauthorized access attempts by individuals outside the firewall are completely owned by the firewall and thus prohibited. The only incoming

068082.0114-052101

traffic should be that which is going to HTTP, HTTPS, and perhaps FTP.

The application server must have an entirely open communication channel to the database servers. The application server will connect to the database server using a single logon account and password. It will open as many connections as necessary (all under this single username and password) and will pool all data requests from all users.

For each user and each session, a special "Security Key" 128 byte encoded string is assigned. Implemented both in the database servers, and in the application servers, this Security Key becomes a time-sensitive passcode that will prove the security authenticity of an incoming request. These security keys can expire after a configurable number of minutes, and they can be assigned only to one user and one session at a time. If a user tries to create two sessions, his first session instantly becomes invalid and no longer usable.

Username and password logons are stored in the database server. The application server fetches the user's input in these fields while logging on and reconciles them against the Logons table in the System Database. If a match is found a Security Key is generated, time-stamped, and linked to the user.

Hacking attempts on a username and password are tracked. For a specific account, sequential invalid logon attempts are counted and recorded. If the bad logon count exceeds the maximum, the account becomes "locked" and only a system administrator can unlock it.

To protect the superuser and admin accounts, these accounts can be restricted to a specific IP address or some other means of machine authentication to ensure that outside hackers have no means to hack into the "root" accounts.

Between the application server and database server is plain text no-encryption. Between the application server and the Internet browsers, there can be either no encryption, or any level of SSL encryption. SSL adds CPU load, but for certain areas of the site might be good to have in place.

#### Operational Scenario

Users will typically be sales representatives whose main objective is to quickly identify high quality leads and determine the reason for such qualification and best method to position their product or service for sale. Users will need to have control over an individual profile, login to their lead site, have a personal workspace which functions as their lead home, view leads on the screen and progressively drill down into: (1) contact information, (2) document summary and (3) original document with highlighted key phrases, perform multi-level searches and sub searches into their lead base by looking at all relevant documents in their set, generate scripted emails or print documents that includes business logic and intelligent extract from the original Internet document, close and rank leads based on subjective criteria, view lead performance reports on those leads within their area, rank leads by time to closure or estimated sale value.

A user session might follow these general steps:  
Login, User completes descriptors, User suggests sources,  
Launch search, Download, Cleanse, Harvest, Highlight,  
Cascade lookup, Prioritize prospects (date, time, rank,  
5 etc), Push to desktop, Web export.

System Platform

INS. 37 Referring again to FIGURES 1 and 2, the server  
functions of system 10 may be partitioned among more than  
10 one piece of equipment. Standard server equipment may be  
used, such as those capable of running Windows 2000  
server software. Other software used to implement the  
invention may include Oracle 8i Enterprise Edition, Cold  
Fusion 4.5 Enterprise for Windows, Verity or Thunderstone  
15 for search engine, and Cognos or Seagate products for  
report generation.

INS. 38 System 10 is based on a client/server architecture.  
The server of system 10 can reside on Windows NT/2000  
Server, Sun Solaris (Unix), and AIX (Unix) architectures.  
20 The client may be any web browser that supports Java 1.1  
(or higher) plug-in, such as Microsoft IE 4.0 (or  
higher), or Netscape Communicator 4.0 (or higher). These  
web browsers run on most major platforms, such as  
Windows95/98/NT/2000, Unix (Sun Solaris, AIX, Linux,  
25 etc), or OS/2 and MacOS.

Other Embodiments

Although the present invention has been described in  
detail, it should be understood that various changes,  
30 substitutions, and alterations can be made hereto without  
departing from the spirit and scope of the invention as  
defined by the appended claims.